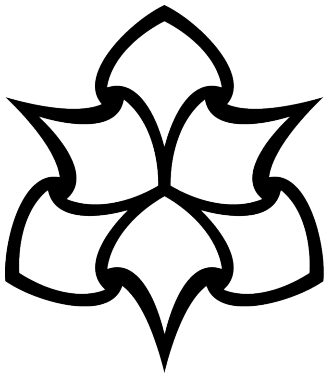

Expert Systems With Applications 127 (2019) 353–369

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach

Lewis Evans[a,*], Majdi Owda[a], Keeley Crockett[a], Ana Fernandez Vilas[b]

[a] School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University M1 5GD UK, Manchester, England United Kingdom
[b] I&C Lab. AtlantTIC Research Centre, University of Vigo, 36310, Pontevedra, Spain

ARTICLE INFO

Article history:
Received 22 November 2018
Revised 11 March 2019
Accepted 12 March 2019
Available online 12 March 2019

Keywords:
Cashtag collision
Twitter
Stock market
Data fusion
Machine learning
Natural language processing

ABSTRACT

Investors utilise social media such as Twitter as a means of sharing news surrounding financials stocks listed on international stock exchanges. Company ticker symbols are used to uniquely identify companies listed on stock exchanges and can be embedded within tweets to create clickable hyperlinks referred to as cashtags, allowing investors to associate their tweets with specific companies. The main limitation is that identical ticker symbols are present on exchanges all over the world, and when searching for such cashtags on Twitter, a stream of tweets is returned which match any company in which the cashtag refers to - we refer to this as a cashtag collision. The presence of colliding cashtags could sow confusion for investors seeking news regarding a specific company. A resolution to this issue would benefit investors who rely on the speediness of tweets for financial information, saving them precious time. We propose a methodology to resolve this problem which combines Natural Language Processing and Data Fusion to construct company-specific corpora to aid in the detection and resolution of colliding cashtags, so that tweets can be classified as being related to a specific stock exchange or not. Supervised machine learning classifiers are trained twice on each tweet – once on a count vectorisation of the tweet text, and again with the assistance of features contained in the company-specific corpora. We validate the cashtag collision methodology by carrying out an experiment involving companies listed on the London Stock Exchange. Results show that several machine learning classifiers benefit from the use of the custom corpora, yielding higher classification accuracy in the prediction and resolution of colliding cashtags.

1. Introduction

Investors make use of many online discussion channels when deciding to make investments on stock markets. Such information is presented within Financial Discussion Boards (FDBs), news corporations (e.g. Financial Times), broker agency websites, and social media platforms. Recently, Twitter has become a popular platform for investors to disseminate stock market information and discussion (Brown, 2012). Many large organisations are also using Twitter as a platform to obtain and share information relating to their products and services (Huizinga, Ayanso, Smoor, & Wronski, 2017).

Companies are identified on stock markets through the use of ticker symbols, which are typically one to four characters in length (depending on the exchange) and are unique to an exchange, e.g. the TSCO ticker refers to Tesco PLC on the London Stock Exchange (LSE). The use of these ticker symbols within tweets on Twitter are referred to as cashtags and allow investors to participate in discussions and view news regarding a specific company at a moment's notice (Rajesh & Gandy, 2016). Cashtags are clickable links embedded within tweets which mimic the company's ticker symbol, prefixed with a dollar-symbol (e.g. $TSCO cashtag on Twitter refers to Tesco PLC) (Oliveira, Cortez, & Areal, 2016). Cashtags were originally introduced by Stocktwits[1] to allow users to link companies with their posts. Twitter introduced the feature of cashtags in 2012 to allow their users to associate specific companies with their tweets (Li, Shah, Nourbakhsh, Fang, & Liu, 2017). A tweet can contain multiple cashtags, with the only limitation being the character limit imposed upon Tweets, which was recently increased to 280 characters.

The main limitation of cashtags is that they are susceptible to colliding with an identical cashtag belonging to a company listed on another exchange, a phenomenon we refer to as a cashtag collision. As tweets are typically short in length, they can be an in-

* Corresponding author.
E-mail addresses: l.evans@mmu.ac.uk (L. Evans), m.owda@mmu.ac.uk (M. Owda), k.crockett@mmu.ac.uk (K. Crockett), avilas@det.uvigo.es (A.F. Vilas).

[1] https://stocktwits.com/.

https://doi.org/10.1016/j.eswa.2019.03.019

dispensable tool for investors to discuss recent events relating to companies. The presence of colliding cashtags, however, can result in investors having to decide if the tweets returned via their cashtag search actually relates to the company in which they are interested in. Investors not aware that Twitter does not distinguish multiple companies over different stock exchanges with identical ticker symbols could have made investments based on information which is not pertinent to the company in which they thought it was. This is even more problematic if investors use automatic analysis tools to measure the popularity of a certain cashtag or other social media metrics.

Throughout this paper we refer to a cashtag collision as one of two scenarios: (1) two identical tickers which refer to different companies (e.g. $TSCO refers to Tesco PLC on the LSE, but also refers to the Tractor Supply Company on the NASDAQ) and (2) two identical tickers which refer to the same company which has multiple listings on different exchanges (e.g. $VOD refers to Vodafone Group PLC on both the LSE and the NASDAQ). We anticipate that the second scenario will be particularly difficult to detect and resolve, as the same company which is listed on multiple exchanges does not have many features which can distinguish them apart (e.g. VOD on both exchanges will have the same company name and CEO).

The issue of colliding ticker symbols is not just isolated to Twitter, several other news websites which depend on the automatic assignment of news articles to specific companies based on their ticker symbols can also suffer from incorrect assignment of news articles. Yahoo! Finance, for example, incorrectly associates Tesco PLC's (LSE) Regulatory News Service (RNS) statements with the Tractor Supply Company (NASDAQ), which could sow confusion for potential investors who depend on such news sources.

This paper introduces a novel methodology for the detection and resolution of colliding cashtags on Twitter.

We train traditional supervised machine learning algorithms twice on each tweet to classify if a tweet relates a specific exchange-listed company or not. One classifier is trained on a sparse vector of the tweet text alone, while a second classifier is trained on both the sparse vector and other features contained within a company-specific corpus. The cashtag collision resolution methodology introduced in this paper is a generalised approach which can be applied to any stock market. We validate the cashtag collision resolution methodology by carrying out an experiment involving companies listed on the LSE (discussed in detail in Section 4).

The main contributions of this paper can therefore be summarised as follows:

- We highlight the prevalence of colliding cashtags on Twitter.
- We define two related methodologies for (1) the fusing of company information to create company-specific corpora, and (2) resolving cashtag collisions through the use of traditional supervised learning classifiers.
- We demonstrate that several of the classifiers see significant performance increases, in respect to a metric used when there is a class imbalance, when assisted by company-specific corpora.

These contributions address a problem which has yet to be discussed within the literature. Several previous works involving the analysis of cashtags could have been susceptible to incorrect analysis and results due to the subtlety of colliding cashtags.

The remainder of this paper is organised as follows: Section 2 introduces the main motivation of this paper, challenges associated with colliding cashtags, and the research questions we aim to answer. Section 3 explores the related work involving cashtags, disambiguation on Twitter, data fusion, and the use of custom corpora. Section 4 provides an overview of an experiment which has been designed to validate the cashtag collision resolution methodology. Section 5 provides an overview of the data used in this experiment. Section 6 introduces the company corpora creation and data fusion methodology. Section 7 provides a high-level exploratory analysis of the data. Section 8 details the cashtag collision resolution methodology for classifying a tweet as belonging to a specific exchange or not. Section 9 discusses the results of the experiment. Section 10 draws a conclusion and proposes future work relating to cashtag collisions.

Table 1
Disparity of ticker symbols (Vodafone PLC).

Exchange	Reuters Instrument Code (RIC)	Bloomberg Ticker	Google Finance Ticker
LSE	VOD.L	VOD:LON	LON: VOD
NASDAQ	VOD.O	VOD:US	NASDAQ:VOD

2. Cashtag collision challenges

This section presents the motivation, challenges and the research questions this paper will answer.

2.1. Motivation

Although the main limitation of cashtags is Twitter's inability to distinguish between identical cashtags which refer to companies listed on different exchanges, it is also important to mention that the structure of ticker symbols differ across the internet. As Twitter does not adopt or enforce a way for users to include the exchange symbol when referring to a company ticker symbol, as other websites do, a methodology for classifying a tweet as belonging to a specific exchange would benefit both individual investors and businesses alike. Currently, tweets need to manually analysed by the human eye to determine what company is being referred to if no exchange-specific information is available in the tweet, wasting precious time.

2.2. Key challenges

The reason that collisions occur on Twitter is that Twitter has yet to formalise or enforce rules relating to embedding cashtags in tweets. Similar to hashtags, users are free to create their own cashtags by simply prefixing any word with a dollar-symbol, meaning no exchange-specific information needs to be present in the tweet for it to be published. When news is published on websites such as Google Finance and Reuters, a pre-determined rule is often adhered to, in that the exchange in which the company sits on is featured in the ticker symbol. Companies are identified on Reuters, Bloomberg, and Google Finance by the formats shown in Table 1, all of which feature the exchange of the company within the ticker symbol.

Another challenge is that some of the more popular ticker symbols (e.g. WEB) can feature on multiple exchanges (Table 2), making it increasingly more difficult for an investor to decipher which company a tweet refers to.

A challenge relating to the application of Natural Language Processing (NLP) to this field is that text classification is often performed on documents which contain a large collection of words to assist a classifier in determining which class a document belongs to. Tweets, however, are limited to only containing a limited number of words due to the character limit (Gerber, 2014), meaning tweets may not feature enough information within them to provide an accurate classification as to whether or not the tweet relates to a specific exchange company. The lack of textual information in tweets can be overcome by creating a custom corpus for

Table 2
Example LSE ticker collisions.

Ticker	LSE Company	Colliding Exchange / Company Name
WEB	Webis Holding PLC	NASDAQ / Web.com Group, Inc EURONEXT / Warehouses ASX / Webject Ltd
MED	Medaphor Group PLC	NYSE / Medifast EURONEXT / Medasys ASX / Merlin Diamonds Ltd
STL	Stilo International PLC	NASDAQ / Sterling Bancorp BSE / STL Global Ltd ASX / Stargroup Limited

each exchange-listed company via data fusion techniques, which can then be consulted to assist in the classification process.

2.3. Research questions

This paper will answer the following research questions, which will be referred to as RQ1 and RQ2 in subsequent sections:

RQ1: can a tweet's text alone be used to classify a tweet as relating to a specific exchange-listed company?
RQ2: can the creation of company-specific corpora, created through data fusion, improve the classifiers' performance?

With the motivation and research questions outlined, in the next section we discuss the work relating to our proposed methodology and the experiment designed to validate it.

3. Related work

To our knowledge, there has been no related work on the identification or resolution of cashtag collisions. There has, however, been extensive work in other areas related to this research, which include experiments involving cashtags (Rajesh & Gandy, 2016; Vilas, Evans, Owda, Redondo, & Crockett, 2017), word disambiguation on Twitter (Spina, Gonzalo, & Amigó, 2013), the fusion of different data sources (Evans, Owda, Crockett, & Vilas, 2018; Khaleghi, Khamis, Karray, & Razavi, 2013), and the use of custom corpora (Ramos Carvalho, Almeida, Henriques, & Varanda, 2015).

3.1. Cashtags

Previous work on the analysis of cashtags is relatively scant within the literature. Existing work has focused on sentiment analysis of tweets which contain cashtags for the purposes of stock market price prediction, analysing the impact of financial events on Twitter, and uncovering spam bots on Twitter (Bartov, Faurel, & Mohanram, 2017).

Rajesh et al. (2016) collected tweets over a two-month period which contained cashtags for Apple Inc. ($AAPL), listed on the NASDAQ, and Johnson and Johnson ($JNJ), listed on the NYSE, for the purpose of stock market price prediction. Tweets containing these cashtags were then divided into two categories – tweets created during the opening and closing times of the exchanges respectively. A Feedforward neural network was then implemented which took the average sentiment scores for tweets within these categories to predict the opening and closing market prices, reporting a high accuracy. The main limitation of this work is that it only took into consideration two companies, both of which sit on different exchanges.

Vilas et al. (2017) analysed the impact of financial events on Twitter. Tweets containing the keyword "tesco", the hashtag #tesco, or the cashtag $TSCO were collected before and after Tesco PLC announced its merger with Booker Group PLC (both LSE companies). Their findings provided promising evidence that Twitter was permeable to financial events by analysing the rapidness in which Twitter was able to respond to financial events.

Cresci et al. (2018) carried out a large-scale analysis on the presence of spam bots on Twitter. They collected over nine million tweets which contained at least one cashtag of a company listed on one of the five main financial markets in the US over a five-month period. They found that large volumes of tweets containing cashtags of low-value stocks also featured cashtags of more popular, high-value stocks, showing that users attempt to use the popularity of high-value cashtags by "piggybacking" onto them and spreading news of unrelated low-value stocks. They also concluded that large spikes were due to mass, synchronised retweets, showing the presence of bots and that an analysis of retweeting users classified over 70% of them as bots.

3.2. Word disambiguation on Twitter

There have been several studies on word disambiguation on Twitter in recent years (Gorrell, Petrak, & Bontcheva, 2015; Inkpen, Liu, Farzindar, Kazemi, & Ghazi, 2017; Spina et al., 2013). Spina et al. (2013) proposed an approach to disambiguating company names which are mentioned in tweets. Their approach relies on positive and negative filter keywords which, when found within the text of a tweet, can help to establish if a tweet refers to a specific company. For example, the term "ipod" is considered a positive filter keyword for the company Apple, whereas the word "crumble" has a negative shift. They identify keywords for specific companies by automatically collecting terms listed on the organisation's Wikipedia page and the company URL and then manually associate positive and negative terms with companies. Tweets classified by such keywords were then used with a supervised machine learning algorithm, obtaining a classification accuracy of 73%. Research which involves the use of performing NLP on tweets often use NLP models which are specially trained on a corpus of tweets (Pinto, Gonçalo Oliveira, Alves, & Oliveira, 2016).

3.3. Data fusion

Data fusion is a well-known technique which can be used to enhance the quality of data (Bentley & Lim, 2017). The fusion of heterogeneous data has been considered for a wide variety of problems, including navigation systems, military, habitat mapping, and the fusion of heterogeneous financial market data (Evans et al., 2018). Data fusion can be a challenging task to undertake for reasons such as disparate and heterogeneous data which cannot easily be combined together, specifically if the fusion needs to be performed over a varied temporal space (Khaleghi et al., 2013).

Bharath Sriram (2010) provides five broad categories of tweets (opinions, private messages, deals, news, and events) for the purpose of improving information filtering (associating tweets with a specific category or topic). They first trained a Naïve Bayes model on a Bag of Words (BoW) alone, and then combine this BoW with other features such as the author name of the tweet and occurrence of user mentions within the tweet. They were able to obtain improved classification accuracy scores when the Naïve Bayes model considered both the BoW and the supplementary features combined, showing that the consideration of supplementary features can be of benefit to a classification task.

3.4. Custom corpora

Several previous works (Cheng & Ho, 2017; Moreno-ortiz & Fernández-cruz, 2015; Ramos Carvalho et al., 2015; Wood, 2015) have utilised custom-made corpora for tasks in which ready-made or "generic" corpora are not sufficient for the task at hand due to domain-specific vocabulary. Ramos Carvalho et al. (2015) proposed

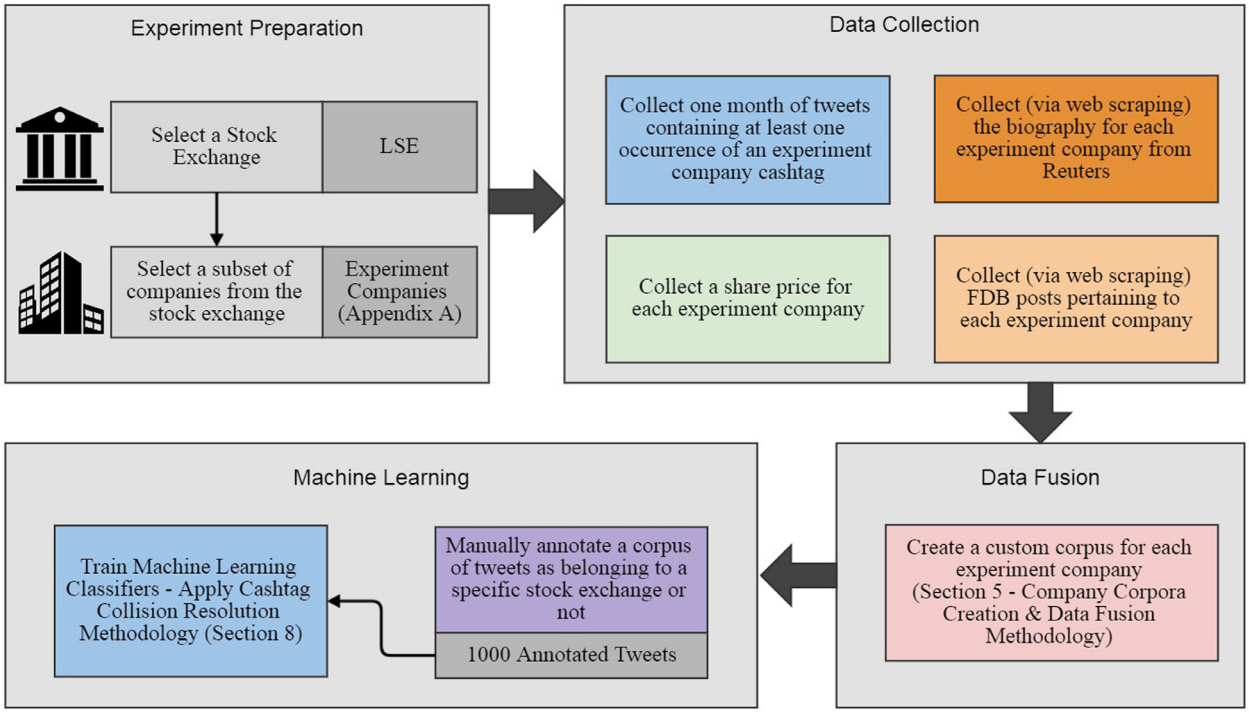

Fig. 1. Experiment overview.

a technique to create domain-specific corpora to convert source code identifiers to their equivalent full name counterparts (e.g. a method named "strcmp" can be split into the words "string, compare"). Their work did note limitations in that, without a domain corpus, translations between source code identifiers to full words can be difficult to achieve.

This paper attempts to address several of the challenges outlined in the related work we have just explored. In regards to cashtag analysis, we consider a larger cashtag space than that explored in (Rajesh & Gandy, 2016) by examining 100 company cashtags. Although we do not attempt to disambiguate between specific keywords found within tweets, we do attempt to disambiguate tweets by classifying tweets as relating to an exchange-listed company or not. In regard to data fusion, we do not attempt to fuse data based on time. Instead, we fuse company-specific information together from three different external data sources in one batch, eliminating the challenges associating with real-time data fusion. This fusion process supports the creation of custom company corpora which will contain information that is specific to each company.

The next section will provide a high-level overview of an experiment to validate the cashtag collision resolution methodology.

4. Experiment details

An experiment (Fig. 1) has been designed which involves creating a custom corpus of company-specific information for 100 preselected companies.

4.1. Experiment preparation

For the purposes of this paper, we validate our cashtag collision resolution methodology by performing an experiment using 100 LSE companies (listed in Appendix A). The LSE has been chosen due to having a popular FDB associated with it which is dedicated to LSE-listed companies, allowing web scraping techniques to yield information specific to companies listed on that exchange. The LSE is formed of two sub-markets; the Alternative Investment Market (AIM) and the Main Market (MM). The AIM is suited for growing businesses and has a more flexible regulatory system than the MM (Barnes, 2017).

4.2. Company selection

In regards to the 100 companies used in our experiment, we select 50 companies from each sub-market (25 of which have a known collision with another company listed on one of the exchanges in Table 3, the remaining 25 with no known collision with the exchanges). Companies are selected randomly from each of the LSE's ten different industries (basic materials, consumer goods, consumer services, financials, health care, industrials, oil & gas, technology, telecommunications, and utilities). Only companies which have been listed on the LSE for at least two years were eligible in this selection process, to ensure that they are well-established and to maximise the chance of collecting tweets containing cashtags relating to LSE-listed companies.

4.2.1. Data collection

In order to ascertain if a tweet relates to a specific exchange-listed company, such as the LSE, data from multiple, reputable sources will be collected and combined to ensure a reliable reference to each of the LSE-listed companies is available.

Tweets pertaining to the 100 experiment companies are collected in real-time via the Twitter Streaming API, which collects no more than 1% of all tweets tweeted in real-time (Abdeen, Wu, Erickson, & Fandy, 2015). Descriptions for each of these companies are web scraped from Reuters so that certain keywords associated with the LSE-listed cashtag company can be obtained, which will be beneficial later to ascertain how many words within the tweets are also found to be in LSE-listed company's biography. FDB posts are then collected from an FDB which is dedicated to LSE companies, allowing us to collect posts which are specific to the LSE companies used in this experiment.

Finally, a share price for the company is collected to assist in the manual annotation of the tweets, this can be a helpful attribute

Table 3
Major stock exchanges (by Market Capitalisation) as of April 2018.

Exchange	Country	Companies Listed	Market Cap (USD bn)	Ticker Style
New York Stock Exchange (NYSE)	United States	3143	21,377	1–9 Characters
NASDAQ	United States	3302	9585	1–6 Characters
Euronext	European Union	923	4388	2–5 Characters
London Stock Exchange (LSE)	United Kingdom	2027	4297	3–4 Characters
Bombay Stock Exchange (BSE)	India	5749	2175	3–11 Characters
Australian Securities Exchange (ASX)	Australia	2255	1428	3 Characters

Table 4
Data sources & collection techniques.

Data Source	Collected Via	Data Collecting	Date(s) Collected
Twitter (Structured)	Tweepy	Any tweets which have at least one occurrence of a cashtag relating to the experiment companies (Appendix A).	16/4/2018–16/5/2018
Financial Discussion Board - London South East (Unstructured)	Scrapy	Post ID Subject Date Share Price (at the time of posting) Opinion Author Number of Posts (of the Author) Premium Member (True/False) Post-Type Text	22/04/17–22/04/18 (1 Year)
Reuters (Unstructured)	BeautifulSoup	Company Name Company Description Company CEO	22/04/18
AlphaVantage (Structured)	AlphaVantage API	Share Price	22/04/18

if a tweet contains a reference to a share price when little other information is available. Section 5 will provide more details on the data collected for this experiment.

4.2.2. Data fusion

The company descriptions, FDB posts, and the company share prices are combined to create a company corpus for each of the experiment companies. These corpora will assist the machine learning classifiers later to establish if there is any correlation between the features present within the tweet and the features present in the associated LSE-company corpus. Section 6 provides a detailed overview of this corpora creation methodology.

4.2.3. Machine learning

Traditional supervised machine learning algorithms are trained twice on each tweet (Section 9.3) to classify if a tweet relates to an LSE-listed company or not. One classifier is trained on a sparse vector of the tweet text alone, while the second classifier is trained on the sparse vector and other features made available from the custom corpora. Section 9 contains more details on the classifiers used for this experiment, including the results obtained. We hypothesise that the classifiers which are trained on the combined features will perform better in respect to the traditional performance metrics (accuracy, precision, recall).

In the next section, we provide an overview of the different data sources used in this experiment, along with the motivation for their use in being fused together to create company-specific corpora.

5. Data sources

We now introduce the data sources, beginning with Twitter, and then the fusion data sources which will be fused together to create company-specific corpora, which will be utilised in Section 6 when the data fusion methodology is introduced. A complete list of the data sources, along with the methods of collection, and dates in which the data is collected, is provided in Table 4.

5.1. Twitter

We only collect tweets which have at least one occurrence of a cashtag belonging to at least one of the experiment companies. In total, we have collected 86,539 tweets, which include tweets having collisions and tweets without. These tweets cover a one-month period from 16/4/2018 to 16/5/2018.

5.2. Fusion data sources

The data sources listed below are used specifically in the fusion process, company-specific information from Reuters, an FDB (specifically for our experiment, London South East), and AlphaVantage will be used to create company-specific corpora. Preprocessing techniques are explained in Section 6, when the data fusion methodology is introduced.

5.2.1. Reuters

The Reuters finance section contains a description for every company listed on all the major stock exchanges around the world. The description typically consists of a brief paragraph which details relevant company information such as the company industry, location of operation, and other pertinent information. Keywords found within the description could help to establish if a tweet relates to an LSE-listed company or not. The description for each company has been scraped via BeautifulSoup,[2] a Python library suitable for scraping websites.

5.2.2. Financial Discussion Board – London South East

A popular FDB used by investors trading on the LSE, London South East features a sub-forum for every company listed on the LSE in which investors can discuss news and events for a specific company. FDB posts can help determine what topics are being discussed by investors in relation to the specific company and its corresponding subforum.

[2] https://www.crummy.com/software/BeautifulSoup/.

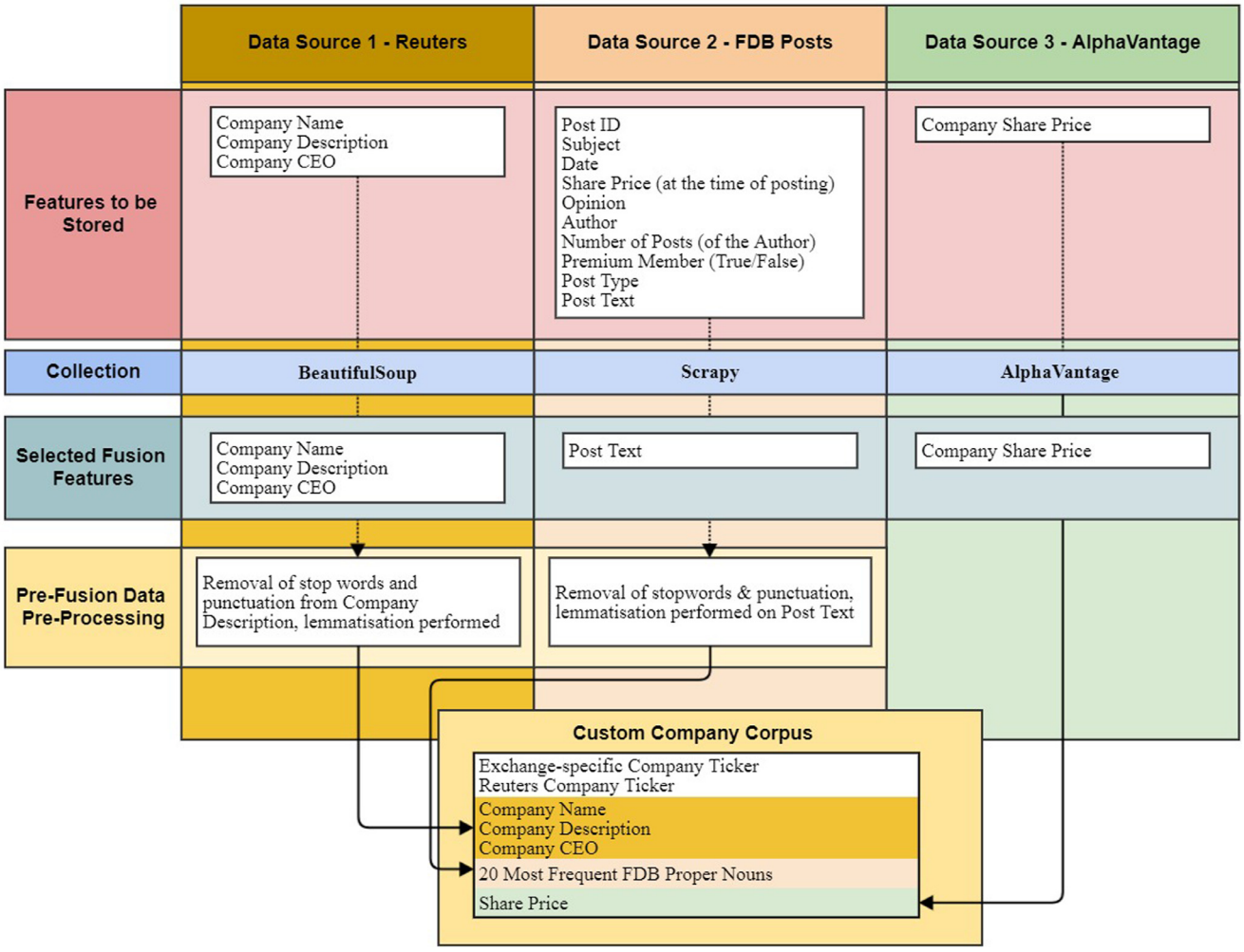

Fig. 2. Custom Corpus Creation through data fusion.

As financial posts span across multiple pages, the open-source web crawling framework, Scrapy,[3] has been used to extract the posts of each of the discussions for the 100 sub-forums. London South East records stock discussion posts going as far back as one year. We have collected all of the posts available for each of the experiment companies.

5.2.3. AlphaVantage

AlphaVantage[4] offers real-time stock market prices for shares listed on stock exchanges. We have collected a recent share price for each of the experiment companies, which may prove to be a valuable source of information if tweets are found to frequently feature share prices, as this could help to distinguish which company is being referred to. Now that the different data sources have been introduced, we now present the methodology for creating individual company corpora through the use of data fusion.

6. Company corpora creation & data fusion methodology

This section will present the methodology (Fig. 2) for creating company-specific corpora through the use of data fusion. We begin by describing the corpora creation steps and exploring the benefits and associated challenges of performing this data fusion on the different data sources.

6.1. Corpora creation

This section will provide more details on the corpora creation methodology, which includes the features from each data source to be collected, the collection method, selected fusion features, and the data pre-processing steps to be carried out on each of the fusion data sources.

6.1.1. Feature selection & collection

The first step of the fusion process is to collect each of the fusion data sources listed in Section 5.2. The Reuters company descriptions for each of the experiment companies have been collected via the BeautifulSoup library. FDB posts have been collected via the Scrapy library, with the share prices being collected using AlphaVantage's API.

6.1.2. Fusion features

Although the Reuters company descriptions and the FDB posts contain several features which are being stored, not all of these features will provide benefits when being contained in a company's corpus.

Table 5 Outlines the features to be fused and contained within a company corpus, along with the reasoning behind these choices.

[3] https://scrapy.org/.
[4] https://www.alphavantage.co/.

Table 5
Corpora data sources fusion features.

Data Source	Fusion Data Features	Reasoning
Reuters	Company Name	The company description is the key feature being extracted from Reuters, keywords found within a tweet which are also contained within the custom corpus can be indicative of a tweet relating to the LSE-listed company.
	Company Description	
	Company CEO	
FDB Posts	Post Text	Although FDB posts contain many features, the most valuable is the textual body within the FDB post. Investors sharing news on FDBs often include other pertinent details such as the company's chief competitors, which can help to establish if a tweet related to the company in question.
AlphaVantage	Share Price	The share price for the company can assist in the manual annotation of the tweet dataset. For each ticker contained within the tweet, the associated ticker company's share price can be extracted from the corpus to assist the annotation process.

Table 6
NER & data pre-processing techniques.

	Data Source	Feature	Named Entity Recognition	Pre-processing Techniques		
				Stop word Removal	Lemmatisation	Other Removal
Fused Data Sources	Twitter	Tweet Text	Proper Nouns (NNP)	√	√	Removal of URLs
	Financial Discussion Board Posts	Post Text		√	√	
	Reuters	Company Description		√		
	AlphaVantage	Share Price	No Pre-processing required			

6.1.3. Data Pre-Processing

An important part of the fusion process is to perform common pre-processing techniques before the fusion process begins. This includes reducing the dimensionality of the data by removing commonly occurring low-value words and transforming them into their non-inflected form. Table 6 summarises the pre-processing and other cleaning techniques performed on each of the data sources.

6.1.3.1. Named Entity Recognition. The lack of context in short queries (i.e. tweets), due to the character restriction, makes the task of recognising entities particularly difficult for full-text off-the-shelf Named Entity Recognition (NER) (Eiselt & Figueroa, 2013). We have utilised NER by selecting the 20 most frequent proper nouns from each of the FDB company sub-forums. A proper noun being defined as "a name used for an individual person, place, or organisation, spelt with an initial capital letter". This allows us to capture names of people and organisations being mentioned in user posts which can then be used later to record the number of LSE-listed company FDB proper nouns present in the tweets.

6.1.3.2. Stop word Removal. The removal of stop words in the tweets, FDB posts, and Reuters company descriptions has been performed using Python's NLTK package,[5] which includes a pre-built corpus of common English stop words which we use to perform stop word removal from each data source.

6.1.3.3. Lemmatisation. The NLTK has also been utilised to perform lemmatisation on the Reuters company descriptions and all of the tweets' text in order to reduce the number of words, allowing us to reduce the sparsity of our bag of words (discussed in Section 8.2.1) (Jivani, 2016).

6.2. Data fusion challenges

One of the key challenges present in this data fusion process is the heterogeneity of the three data sources. Reuters descriptions are static in the nature that this description will likely stay the same for years. FDB posts are dynamic in the sense that investors will likely be discussing recent news and events relating to a specific company.

As our approach relies on freely-available public data sources, there is the added risk that any of these data sources could suddenly become unavailable, meaning alternative features from other sources may need to be relied upon. Web scraping techniques in particular are susceptible to failing should the structure of a web page change. Utilising services which provide structured data, such as AlphaVantage, also run the risk of service shortages or their associated APIs becoming unavailable or deprecated.

Each of the data sources considered for this experiment do have reliable alternatives. Descriptions for companies can also be obtained from other reputable financial market news providers, such as Bloomberg. There are also other FDBs which do focus specifically on the LSE, although the structure for scraping posts from this FDB is significantly more challenging due to the way the websites structures its web pages. Share prices from AlphaVantage could also be obtained from web scraping, although share prices obtained in this way would likely be outdated when compared to real-time market prices.

In the next section, we perform a high-level exploratory data analysis of the collected data in order to better understand the nuances of the dataset of tweets and FDB posts.

7. Exploratory data analysis

This section will present a high-level overview of the Twitter and London South East datasets. This analysis is based on all of the tweets and FDB posts gathered for the experiment companies (Appendix A). The goal of this exploratory data analysis is to gain a better understanding of the scale of cashtag collisions, in addition to identifying any particular nuances present in the dataset which may be of importance in the annotation process (Section 8.1).

7.1. Twitter

We begin by exploring the Twitter dataset with an exploration of the cashtags within the tweets. A total of 86,539 Tweets have been collected over a one-month period from 16th April 2018 to 16th May 2018.

Taking into account the full twitter dataset of 86,539 tweets, we begin the analysis by checking how many tweets contain a cashtag which collide with one of the exchanges in Table 3. In total, 55,543 (64.2%) contain a colliding cashtag (based on our definition

[5] https://www.nltk.org/.

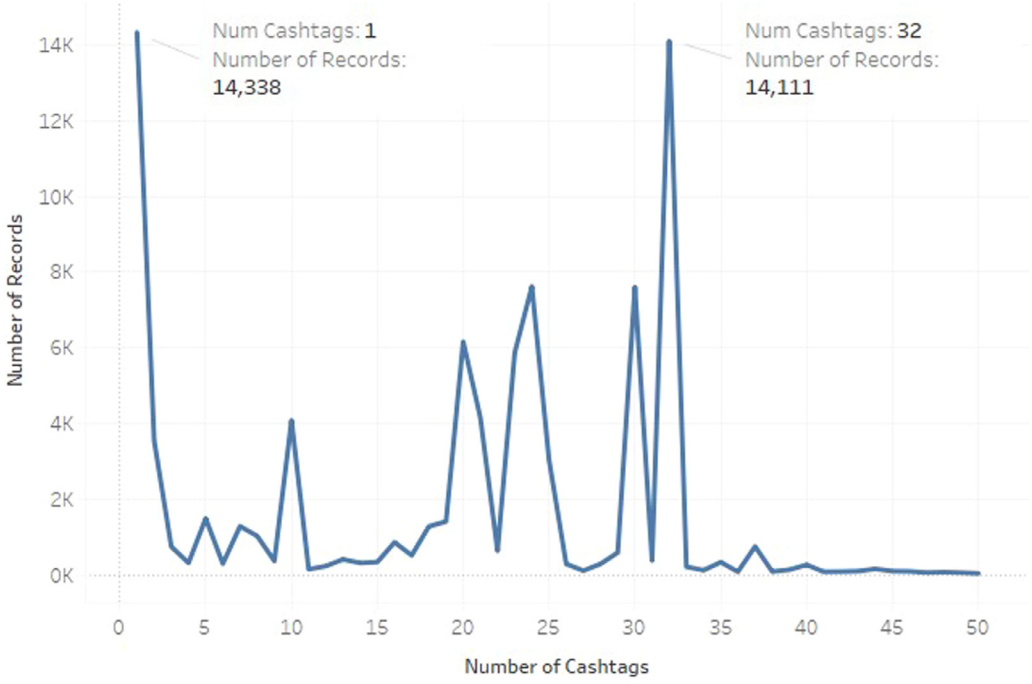

Fig. 3. Cashtag distribution.

in Section 1). This highlights the scale of the problem, which this research is attempting to address.

7.1.1. Cashtag distribution

The number of cashtags present within the tweets in our dataset falls between 1 and 50 (Fig. 3), with significant hikes at 10, 20, 24, 30, and a dramatic increase at 32 which almost exceeds that of tweets containing a single cashtag.

It is a reasonable assumption that the majority of tweets should contain one cashtag, as tweets are limited to 280 characters, allowing only a limited amount of information to be shared. There is no immediate indication as to why there is such a surge of tweets containing 32 cashtags.

7.1.2. Irregular cashtag – BTG

The most dominant cashtag in our dataset is $BTG (Fig. 4), present 58,733 times (tweets can contain duplicate cashtags). A large portion of these BTG tweets (13,309) contain the exact same textual content when not considering hyperlinks embedded within them (Fig. 5), indicating the presence of tweets created by bots. All of these tweets contain 32 cashtags, which explains the hike of cashtag distribution in Fig. 3.

The most frequent word found in BTG tweets ("binance") refers to Binance Coin, a cryptocurrency which is currently ranked in the top twenty of all cryptocurrencies in terms of market capitalisation. There are currently over 1600 cryptocurrencies according to CoinMarketCap,[6] all of which feature their own symbol which can be converted into a cashtag on Twitter, similar to stock market ticker symbols.

The Twitter streaming API provides a structured JSON object for each tweet which contains details relating to the tweet, author, location, amongst other items. A useful attribute for detecting how a tweet was published to Twitter is the *source* field, which provides the medium used to publish a tweet.

A breakdown the most popular Tweet sources in our dataset (Fig. 6) shows a clear presence of unofficial apps generating tweets. We can now therefore conclude that the popularity of BTG cashtag in our dataset is due to the prevalence of automated cryptocurrency bots on Twitter, and that other cashtags may also be susceptible to such noise.

As a substantial number of tweets come from automated bots, this leads to a considerable amount of noise in our dataset. We do not remove these tweets from our dataset, as these tweets are clearly not related to any specific exchange, meaning the word patterns used can be of use when attempting to classify a tweet as being related to a specific exchange or not.

7.2. Financial Discussion Board (London South East) posts

Analysis of London South East company forums is significantly easier to undertake when compared to tweets, as each sub-forum is dedicated to a particular company listed on the LSE, meaning investors choose a sub-forum to discuss a specific company, thus collisions cannot exist in this domain.

7.2.1. Sector posts

The average number of posts per user of the experiment companies (Fig. 7) shows that companies listed on the AIM feature more active discussions across most sectors than their MM counterparts.

Armed with a better understanding of the Twitter and London South East datasets, the next section will introduce the methodology of resolving cashtag collisions.

8. Cashtag collision resolution methodology

The methodology of determining if a tweet contains a colliding cashtag (Fig. 8) involves the vectorisation of the tweet text into a sparse vector (Feature 1 – F1) and combining other supplementary features such as the number of exchange-specific (F2) & non-exchange-specific cashtags (F3), the count of Reuters company description words (F4), and FDB words (F5) found within the tweet so that traditional machine learning classifiers can make correlations between these features. We now proceed with the different

[6] https://coinmarketcap.com/.

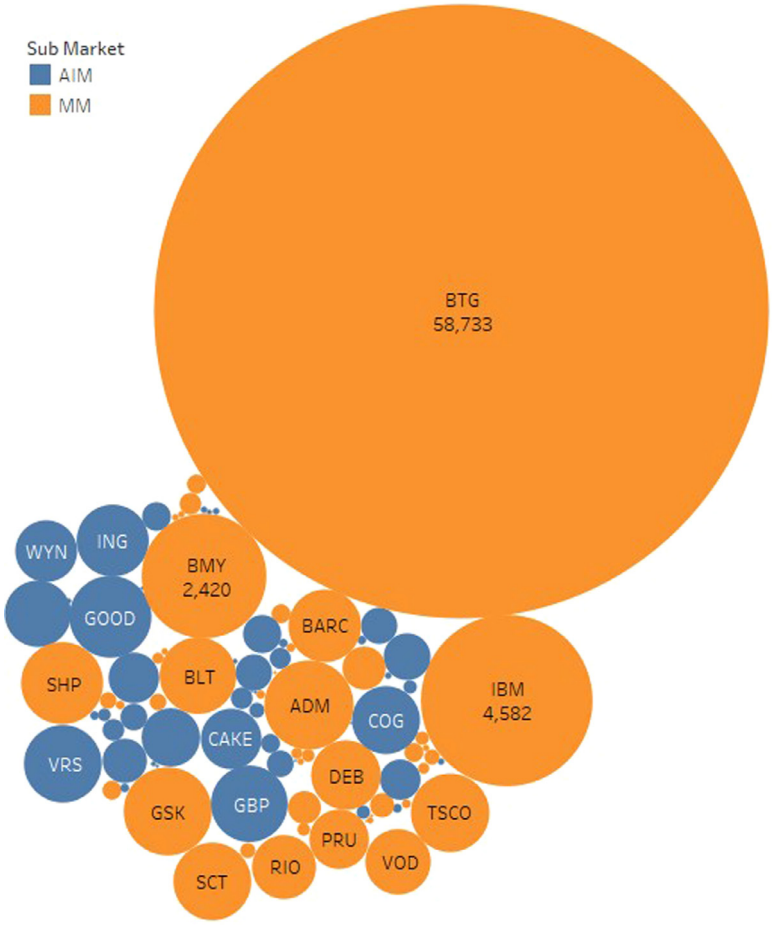

Fig. 4. BTG cashtag dominance.

Fig. 5. Suspected bot tweet.

steps in which we detect and resolve a cashtag collision, beginning with an explanation of our annotated tweet dataset.

8.1. Annotated tweet dataset

In order to answer RQ1&2 (Section 2.3), a labelled dataset of tweets must be created in order to assess the predictive power of the different machine learning classifiers to be trained in Section 9.3. As the cost of creating a manually labelled dataset is time-consuming, particularly when the labelling requires the inspection of each tweet's text and author details, we have manually annotated 1000 tweets with the labels listed in Table 7. Although this is a laborious task even for a relatively small corpus of tweets, this is consistent with previous works relating to tweet annotation (Matsuda, Sasaki, Okazaki, & Inui, 2017; Tjong Kim Sang & van den Bosch, 2013). As the exploratory data analysis showed a heavy presence of cryptocurrency-related tweets, we use three labels to annotate our dataset. A label of zero (0) indicates the tweet does relate to a stock exchange, but not directly to the LSE. A label of one (1) indicates that the tweet directly relates to a company listed on the LSE. A label of two (2) indicates that the tweet references cryptocurrency. In order to ensure consistency in this annotation process, and to ensure high-quality labels (Abraham et al., 2016) are generated, all of these tweets have been manually annotated by a single individual experienced with annotating tweets.

8.1.1. Tweet selection

As evident from the exploratory analysis of the tweets in Section 6, the sheer dominance of the BTG cashtag means that any random selection of tweets will favour tweets containing the BTG cashtag, meaning the classifiers would generalise towards cryptocurrency tweets. To ensure fairness when selecting the 1000 tweets, we first attempt to collect ten tweets for every experiment company ticker (Appendix A). This provided 767 tweets (as some company tickers are not as actively used in tweets compared to others), for the remainder, we collect a random sample of tweets over the one-month time period for a total of 1000 tweets.

8.2. Steps 1–3: Feature design choices

We now provide a motivation for the features used to train the classifiers. Beginning with the sparse vector to represent the text of each tweet.

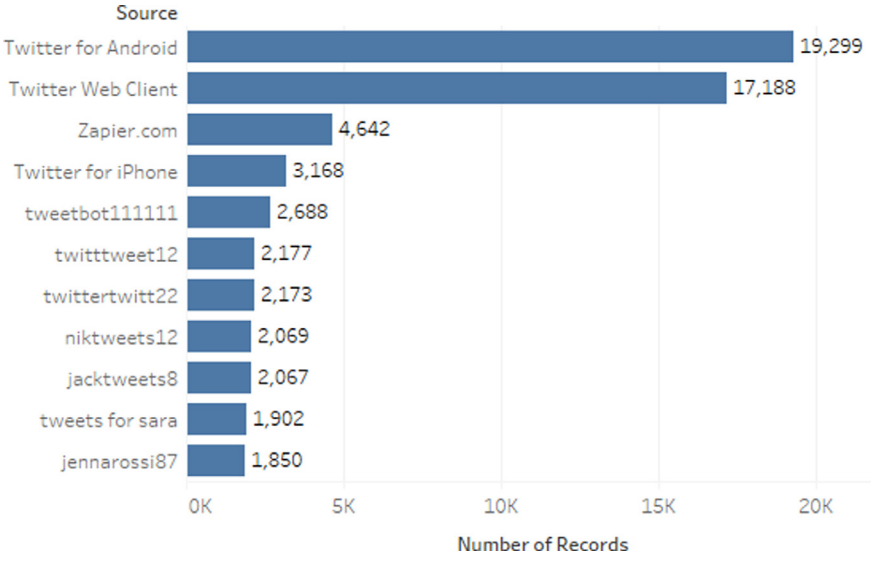

Fig. 6. Tweet sources.

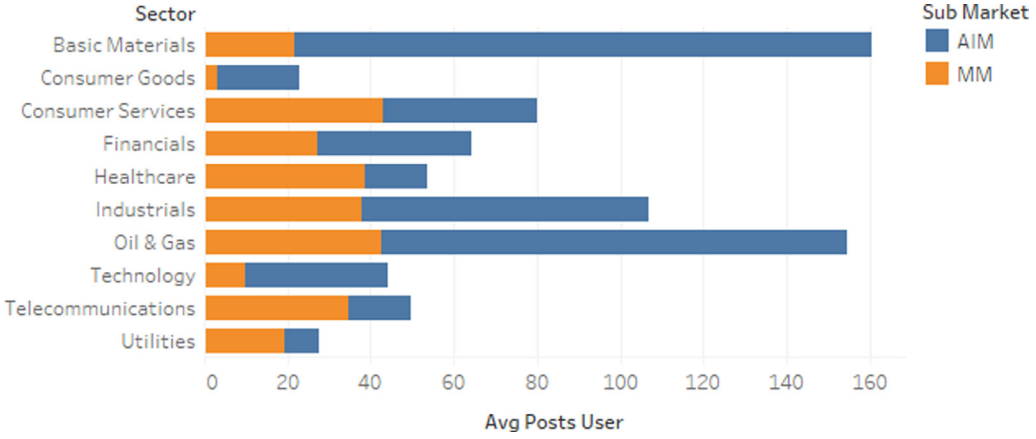

Fig. 7. Average number of posts per user (by sector).

Table 7
Annotated tweet examples.

Label	Tweet Type	Example Tweet
0	Non-LSE related	Cabot Oil & Gas Co. $COG Forecasted to Earn Q1 2018 Earnings of $0.32 Per Share
1	LSE related	Game Digital PLC 55.7% Potential Upside Indicated by Liberum Capital - - $GMD
2	Cryptocurrency related	Sign Up And Recieve 5 (LEGIT) Legitcoin tokens ($10) will be $350 $BTG $ETH $LTC $NXC 2026

8.2.1. Feature 1 (F1) – Sparse vector of tweet text

The first stage of our proposed methodology involves the conversion of all of the tweet text into a sparse matrix. After the removal of stop words and performing lemmatisation, the dimension of our sparse matrix is 1000 × 1860. This sparse matrix is featured in the training of both classifiers. As the cashtags themselves are treated as words, the classifiers will be able to make correlations between the different kinds of cashtags present within a tweet.

In regard to performing such NLP tasks on tweets in preparation for the machine learning classifiers, we elected to use the more general Python NLTK to perform this task. Although Twitter NLP-trained models do exist, none of these models have been trained to deal with the nuances present in our dataset. Although the related research (Pinto et al., 2016) surrounding NLP on tweets found that the performance of standard toolkits (such as NLTK) do not perform as well as Twitter NLP-trained models, this research did not take into account tweets relating to stock discussion, where low-character words such as stock symbols and floating-point numbers are particularly prevalent.

8.2.2. Features 2 & 3 (F2 & F3) – Count of LSE & Non-LSE cashtags in tweet

The number of exchange & non-exchange cashtags present within a tweet can be a strong indication as to whether that tweet relates to a company listed on a given exchange. If a tweet contains one cashtag which relates to the LSE, but also contains a large amount of other cashtags not listed on the LSE, this will undoubtedly assist the classification of such a tweet as being non-LSE related. As all of our tweets contain at least one LSE cashtag, the count of LSE cashtags will always be a minimum of one. As is evident from the exploratory analysis in the preceding section, cryptocurrency tweets have a substantially higher count of cashtags in them.

We have downloaded a list of all ticker symbols relating to the experiment companies listed in Table 3. We then cross-check each

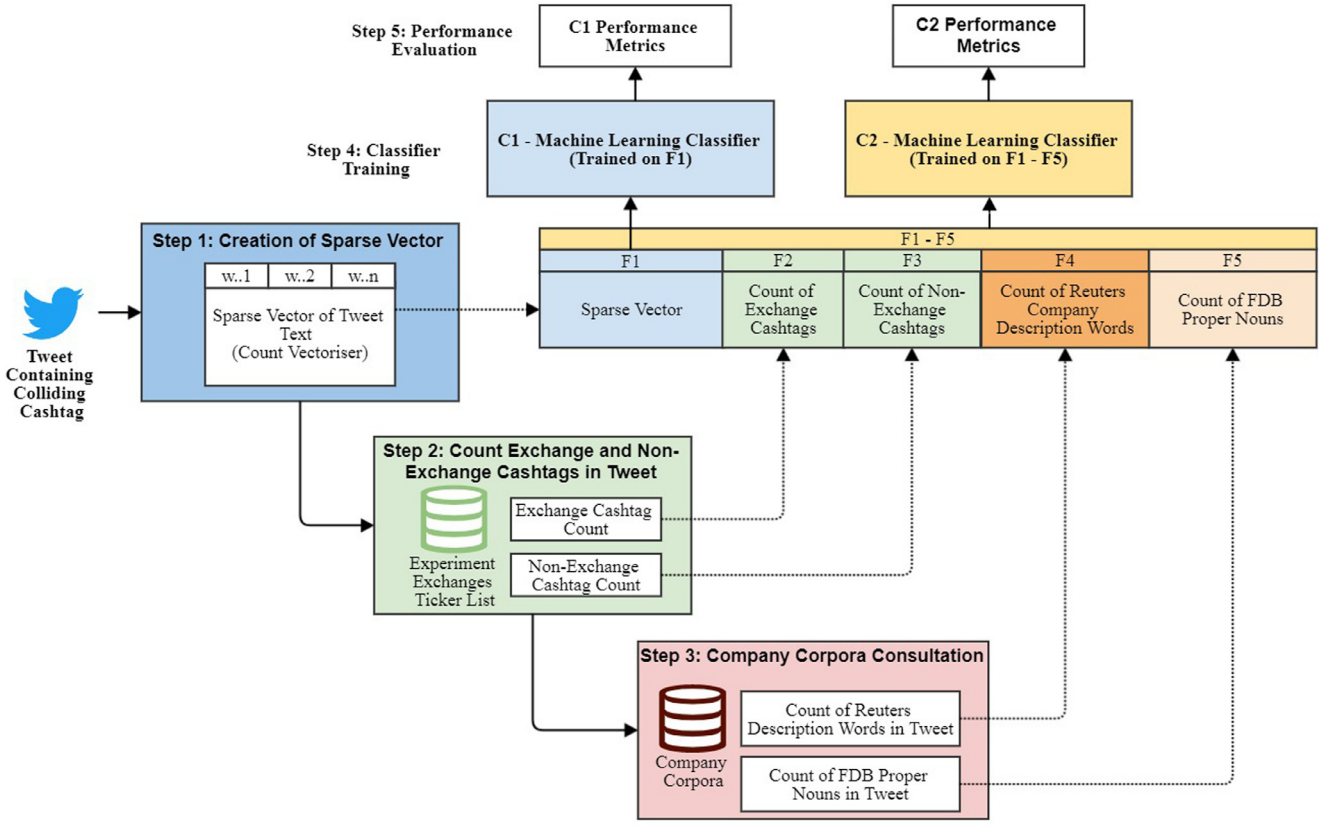

Fig. 8. Cashtag collision resolution methodology.

				Exchange Cashtags in Tweet	Non-Exchange Cashtags in Tweet	Count of Reuters Words in Tweet	Count of FDB Words in Tweet
0	1	..	0	3	29	1	0
0	0	..	0	1	0	3	4
..	..	..	..	..	..	..	..
0	0	..	1	5	2	5	4

Fig. 9. Final sparse matrix representation.

tweet to see how many cashtags within the tweet relate to an LSE-listed company, with the remainder of cashtags being non-LSE cashtags.

8.2.3. Feature 4 (F4) – Count of Reuters description keywords in tweet

The count of words in the tweets which also feature in the tweet's corresponding company corpus can provide strong evidence that a tweet relates to the LSE-listed company. As low-value words have been removed from the description prior to being stored within a company's corpus, words found within the tweet text which also feature in the company description can provide a high correlation that the LSE-listed company is being referenced in the tweet. The LON:TSCO corpus, for example, features words which are able to distinguish it from its colliding company on the NASDAQ, such as "food", "retail", and "united kingdom", which would not be commonly found in tweets referencing the Tractor Supply Company.

Naturally, if two or more companies with a colliding cashtag belong to a similar sector, then this feature of counting the number of word occurrences will not provide as much value. For example, LSE:ABC (Abcam PLC) and NYSE:ABC (AmerisourceBergen Corporation) are both in the Healthcare sector, meaning their respective Reuters biographies will contain similar terminology. To alleviate this, a feature which relies on user-generated terms could be of use, this is our motivation for our final feature.

8.2.4. Feature (F5) – Count of FDB proper nouns in tweet

The final feature we have proposed is to use the most frequent proper nouns found within the FDB posts for each of the LSE-listed companies. The number of FDB proper nouns contained within the tweets could be a helpful indication to establish if a tweet refers to a specific exchange-listed company or not. The sub-forum for Tesco (LSE), for example, has frequently-discussed proper nouns such as Lidl and Aldi – Tesco's chief competitors, allowing a further distinction between LON:TSCO and NASDAQ:TSCO. This feature will be particularly more helpful to solve the more complex collisions

in which two or more companies with the same ticker have the same company name but are listed on different exchanges.

In respect to these five features, we believe that, when combined (Fig. 9), they provide a more robust approach to detect a colliding cashtag tweet, versus using any single feature in isolation.

8.3. Step 4: Classifier training

After a tweet has been represented numerically by transforming it into a sparse vector, and the count of LSE, Non-LSE, Reuters, and FDB keywords have been recorded, this can then be used to train the classifiers. Based on previous works which have seen varying levels of success (Verma Scholar, Professor, & Sofat, 2014), we have chosen to train Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest classifiers. These are each discussed in Section 9. Each of the aforementioned classifiers is trained and tested twice independently. The first classifier (C1) is trained on just the sparse vector of the tweet text (F1) alone, and the second classifier (C2) is trained on the sparse vector and other supplementary features (F1–F5) contained within the company corpora.

8.4. Step 5: Performance evaluation

The final stage of our proposed methodology involves comparing each of the classifiers to determine if a classifier benefits from being trained on the additional features. We compare the performance between the classifiers using the Matthews Correlation Coefficient score, a metric used to assess the performance of a binary classifier which has a class imbalance, discussed in further detail in Section 9.2.

The next section contains the results and discussion of the experiment results.

9. Results and discussion

This section will explore if the consideration of additional features improves the classification performance over the traditional approach of using a sparse vector alone.

The classification of tweets in this experiment is a binary classification problem – a tweet either relates to the LSE (1), or it does not (0). All of the cryptocurrency tweets (labelled 2) have been labelled zero for the training of all of the classifiers. This section will introduce a number of suitable supervised machine learning classifiers, along with their respective benefits, drawbacks, and performance on the annotated dataset.

9.1. Accuracy paradox

Before delving into each of the classifiers used in this experiment, it is important to note why we do not blindly depend on the accuracy of the models as an indication of their respective performance. High accuracy scores can often be misleading as to the predictive power of a classifier. A binary classification problem which features a dominant label can often lead to a misleading accuracy score. In our labelled dataset of 1000 tweets, 642 tweets do not correspond to the LSE, hence being labelled zero. This means if we choose to abandon our machine learning models and predict zero every time, we would achieve a 64% accuracy for free, giving a false indication of predictive power, referred to as the accuracy paradox (Valverde-albacete & Pela, 2014).

9.2. Matthews Correlation Coefficient

A more practical approach to evaluating the results of a binary classifier in which there is class imbalance is the Matthews Correlation Coefficient (MCC) (Boughorbel, Jarray, & El-Anbari, 2017).

Table 8
Logistic Regression results.

	Sparse Vector		Combined Features	
CM	616	26	618	24
	50	308	40	318
MCC Score	0.83		0.86	

Table 9
kNN results.

	Sparse Vector		Combined Features	
CM	609	33	588	54
	73	285	58	300
MCC Score	0.77		0.76	

The MCC score (Eq. (1)) is calculated by using the Confusion Matrix (CM) results using the equation below (where TP = true positive, TN = true negative, FP = false positive, and FN = false negative):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

The MCC score returns a value from −1 to +1. A value of +1 indicates the model makes perfect predictions, 0 indicates the model is no better than random chance, with −1 representing the classifier has made incorrect predictions across the board (Liu, Cheng, Yan, Wu, & Chen, 2015).

Once each of the classifiers' performance has been discussed, we compare the two best performing classifiers (in respect to their MCC score), to determine if the results between the two best performers are statistically significant. H_0 denotes the null hypothesis, which we will attempt to reject at a significance level of five percent. H_1 denotes the alternative hypothesis, which we will attempt to lend support to if we are able to reject H_0.

$$H_0 : MCC_{c_1} < MCC_{c_2}$$
$$H_1 : MCC_{c_1} \geq MCC_{c_2}$$

9.3. Machine learning classifiers

All of the classifiers have been implemented using the skikit-learn library within Python. Each classification model has differing hyperparameters which can affect the performance metrics of the classifier, we find optimal hyperparameters for each classifier through the use of a grid search, which explores a user-specified parameter space to determine the most efficient combination of hyperparameters in respect to a scoring metric (we elect to choose the best hyperparameter combinations based on the MCC score) (Öğüt, Mete Doğanay, & Aktaş, 2009). A common approach suggested by Geron (2017) is to start with a coarse grid search covering a wide parameter space, and then a finer grid search based on the best values found – we have adopted this approach. Internal 10k-fold cross validation has been used for each classifier using an 80/20 train/test split.

A complete table of results for each classifier is provided in Table 14.

9.3.1. Logistic Regression

The first classifier we consider is Logistic Regression (LR), due to its suitability for relatively small training sets (Perlich, Provost, Simonoff, & Stern, 2003). The LR results (Table 8) show an observable increase in the MCC score when the classifier is trained on the combined features when compared to just the sparse vector alone.

9.3.2. K-Nearest Neighbours

The next classifier trained is the K-Nearest Neighbours (kNN) classifier. The kNN results (Table 9) show that the classifier trained on the combined features does not yield a better MCC score compared to the sparse vector alone.

Table 10
SVM results.

	Sparse Vector		Combined Features	
CM	614	28	624	18
	42	316	33	325
MCC Score	0.85		0.89	

Table 11
Naive Bayes results.

	Sparse Vector		Combined Features	
CM	556	86	555	87
	20	338	14	344
MCC Score	0.79		0.80	

Table 12
DT results.

	Sparse Vector		Combined Features	
CM	593	49	604	38
	61	297	66	292
MCC Score	0.76		0.77	

Table 13
RF results.

	Sparse Vector		Combined Features	
CM	620	22	622	20
	63	295	65	293
MCC Score	0.81		0.81	

9.3.3. Support Vector Machine

SVMs have had successful applications in fields such as text classification, handwritten digit recognition, and object recognition (Tong & Koller, 2001). The results of the SVM classifiers are reported in Table 10.

The SVM has outperformed kNN by a wide margin and has also significantly outperformed LR. The SVM trained on the combined features is the top-performing classifier so far.

9.3.4. Naïve Bayes

Next, a Multinomial classifier has been trained, due to its suitability with text classification tasks (Tripathy & Rath, 2017), with the results reported in Table 11.

Although the Naive Bayes has outperformed kNN, it still trails behind LR and SVM.

9.3.5. Decision Tree

The Decision Tree (DT) results (Table 12) show that there is a minimal difference between both classifiers, with the classifier trained on the combined features marginally ahead in terms of the MCC score.

9.3.6. Random Forest

Random Forest (RF) classifiers have become increasingly popular, due to being more robust to noise than single classifiers (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012). The RF classifier results (Table 13) perform almost identical, suggesting that the consideration of combined features does not impact the performance of the RF classifier.

9.4. Discussion of results

Our preliminary results show that the top performing classifiers, in respect to their MCC score, are LR and SVM, both of which perform significantly better when considering additional features granted by the company corpora. kNN and DT perform slightly worse when considering features present in the company corpora.

Table 14
Classification results.
(F1 = Sparse vector of tweet text, F1-5 = Sparse vector & supplementary/combined features). Metrics (accuracy, precision, recall and f1-score are an average of 10-fold cross-validation).

Algorithm	Features	Accuracy	Precision	Recall	F1-Score	MCC
LR	F1	92.4%	92.2%	86.0%	89.1%	0.83
	F1–F5	93.6%	93.0%	88.8%	90.9%	0.86
kNN	F1	89.4%	89.6%	79.6%	84.6%	0.77
	F1–F5	88.8%	84.7%	83.8%	84.3%	0.76
SVM	F1	93.0%	91.9%	88.3%	90.1%	0.85
	F1–F5	94.9%	94.8%	90.8%	92.8%	0.89
NB	F1	89.4%	79.7%	94.4%	87.1%	0.79
	F1–F5	89.9%	79.8%	96.1%	88.0%	0.80
DT	F1	89.0%	85.8%	83.0%	84.4%	0.76
	F1–F5	89.6%	88.5%	81.6%	85.0%	0.77
RF	F1	91.5%	93.1%	82.4%	87.7%	0.81
	F1–F5	91.5%	93.6%	81.8%	87.7%	0.81

Table 15
McNemar's test results (LR vs SVM).

LR F1-F5 Predictions	SVM F1-F5 Predictions	
	0	1
0	680	40
1	5	275

The experiment results have concluded that **RQ1** (can a tweet's text alone be used to classify a tweet as belonging to an LSE-listed company?) is a resounding yes. All classifiers trained have yielded a respectable performance, not only in terms of the traditional metrics such as accuracy, precision, and recall, but also in respect to their MCC score. In regard to **RQ2** (can the creation of company-specific corpora, created through data fusion, improve the classifiers' performance?), this is dependent on the classifier in question. LR and SVM both perform significantly better when trained on both the sparse vector and addition features granted by the data fusion process.

We can now examine whether the results between LR and SVM are statistically significant in terms of their respective performances between their two classifiers (sparse vector vs. combined features).

9.5. LR vs. SVM

As evident from the initial experiment results, LR and SVM appear to be the best performing classifiers when trained on the combined features. To test if the results are statistically significant, we perform the non-parametric McNemar's test, proposed by (Dietterich, 1998), to test our hypotheses. The McNemar's test is a statistical test used to compare two paired samples when the data are nominal and dichotomous (Mccrum-gardner, 2008).

The p-value result of performing a McNemar's test on the contingency table below (Table 15) is calculated at 0.016. This indicates that the performance between the two classifiers, in respect to when they both predict either 0 or 1, is significantly different to each other. As we know the MCC score for SVM is slightly higher than LR, we can conclude that SVM is the best performing classifier for detecting a colliding cashtag tweet.

9.6. Implementation of cashtag collision

The methodology to detect a colliding cashtag presented in this paper has involved the manual annotation of tweets as belonging to a specific exchange (1) or not (0). A company or investor wishing to use this technique could do so with relative ease by collecting data from multiple data sources to assist in the classification process. As we have only collected tweets from a specific list of 100 company ticker symbols, the classifiers presented in this pa-

per have been generalised to tweets containing such cashtags. This means that any classifier needs to go through a re-training process whenever a new company ticker symbol is introduced on the exchange a company/investor wishes to detect collisions on. Such annotation should be performed by an expert who is able to distinguish between an exchange-specific tweet and a tweet which does not contain exchange-specific information.

10. Conclusion & future work

Prior to this experiment, the scale of colliding cashtags was relatively unknown. We have highlighted that a small sample of just 100 ticker symbols contain a large collision space in Twitter. We have also demonstrated that cashtag collisions are not just isolated to companies listed on stock exchanges but are also impacted by the increasingly dominant cryptocurrency tickers. We have also shown that although the classification of a tweet belonging to a specific exchange can be achieved using the tweet text alone, significant increases in a classifier's MCC score, particularly LR and SVM, can be achieved by providing supplementary features to the classifiers.

The novelty of this experiment lies in the feature design choices of the machine learning classifiers. Each of the features benefits the classification task in different ways. The count of Reuters keywords embedded in a tweet can assist in the resolution of the first type of collision outlined in Section 1 (two or more companies with the same ticker, but different company names). The second type of collision (two or more companies with the same ticker, and the same company name), is benefitted from the number of FDB proper nouns found within the tweet, as FDB posts are user-created and reflect recent news and discussion surrounding a specific company. Although the NLP pre-processing techniques used in our experiment have enabled the training of robust classifiers, other NLP techniques used on the various data sources could also have a positive influence on the performance metrics of the classifiers. There may also be other features which can further benefit the classifiers' performance, such as scraping recent news article titles for relevant company keywords and storing such keywords within the company corpora and making use of these when training future classifiers. The supplementary features used to train the second set of classifiers could also provide different degrees of informative power – the count of FDB proper nouns found within the tweet could be of greater benefit than the count of Reuters keywords. Further work in this regard could include quantitative analysis on each of the features to assess how each of these features in isolation benefits the classifiers' performance.

Ideally, a universally-agreed method for referring to a company through the use of its exchange and company ticker should be adhered to. Although Twitter has yet to address this – since cashtags function identical to hashtags, in that users are free to create their own. Our results have shown that this issue is problematic in the sense that 64.2% of tweets collected over a one-month period contained at least one colliding cashtag. As previously stated, the current implementation of cashtags on Twitter can sow confusion for investors who are not aware of the problem of colliding cashtags. The proposed cashtag collision methodology presented in this paper can positively impact businesses and investors by deciding if a tweet relates to a specific exchange or not. The proposed methodology can save businesses and investors precious time by eliminating the need to manually examine tweets for relevant keywords.

The solution to the cashtag collision problem presented in this paper will be utilised in the future by an ecosystem which will aim to monitor multiple communication channels for irregular behaviour relating to stock discussions.

Credit author statement

Lewis Evans: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Resources, Data curation, Writing – Original draft, Writing – Review & editing, Visualization,

Majdi Owda: Conceptualization, Methodology, Validation, Resources, Writing – Review & editing, Supervision, Project Administration, Funding acquisition

Keeley Crockett: Conceptualization, Methodology, Validation, Resources, Writing – Review & editing, Supervision, Project Administration

Ana Fernandez Vilas: Conceptualization, Methodology, Validation, Writing – Review & editing, Supervision, Project administration

Appendix A. 100 LSE companies

Table A.1, A.2, A.3, A.4

Table A.1
Alternative Investment Market (AIM) companies (with collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
88E	88 Energy Limited	Oil & Gas	0	51,693
ABC	Abcam PLC	Health Care	1221	9
ARL	Atlantis Resources Limited	Oil & Gas	69	194
ASC	ASOS PLC	Consumer Servies	229	58
AVN	Avanti Communications Group PLC	Telecommunications	10	1871
BKY	Berkeley Energia Limited	Basic Materials	75	1989
CAKE	Patisserie Holdings PLC	Consumer Services	574	60
COG	Cambridge Cognition Holdings PLC	Health Care	722	14
EMAN	Everyman Media Group PLC	Consumer Services	104	7
EYE	Eagle Eye Solutions Group PLC	Technology	207	7
FLOW	Flowgroup PLC	Industrials	344	8857
GBP	Global Petroleum Limited	Oil & Gas	915	2969
GGP	Greatland Gold PLC	Basic Materials	400	60,023
GOOD	Good Energy Group PLC	Utilities	1034	4
HRN	Hornby PLC	Consumer Goods	1	17
HUNT	Hunters Property PLC	Financials	7	2
ING	Ingenta PLC	Technology	810	0
INSE	Inspired Energy PLC	Industrials	129	194
MTR	Metal Tiger PLC	Financials	112	6747
MUL	Mulberry Group PLC	Consumer Goods	3	0
NAK	Nakama Group PLC	Industrials	308	8
PLUS	Plus500 Ltd	Financials	256	216
TRB	Tribal Group PLC	Technology	8	3
VRS	Versarien PLC	Basic Materials	941	4642
WYN	Wynnstay Group PLC	Consumer Goods	597	2

Table A.2
Alternative Investment Market (AIM) companies (without collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
BGO	Bango PLC	Technology	3	593
BIOM	Biome Technologies PLC	Basic Materials	1	86
BLV	Belvoir Lettings PLC	Financials	4	5
BOO	Boohoo.Com PLC	Consumer Services	39	7012
CLIN	Clinigen Group PLC	Health Care	534	160
CLON	Clontarf Energy PLC	Oil & Gas	58	1532
CRPR	Cropper (James) PLC	Basic Materials	1	9
DX.	Dx (Group) PLC	Industrials	0	732
FEVR	Fevertree Drinks PLC	Consumer Goods	9	729
HZD	Horizon Discovery Group PLC	Health Care	31	16
IMTK	Imaginatik PLC	Technology	2	64
ITQ	Interquest Group PLC	Industrials		28
KOOV	Koovs PLC	Consumer Services	7	1065
LCG	London Capital Group Holdings PLC	Financials	0	442
LWRF	Lightwaverf PLC	Consumer Goods	4	433
MANX	Manx Telecom PLC	Telecommunications	6	9
MYT	Mytrah Energy Limited	Utilities	4	159
NAUT	Nautilus Marine Services PLC	Oil & Gas	74	9
PREM	Premier African Minerals Limited	Basic Materials	29	57,895
SOU	Sound Energy PLC	Oil & Gas	26	40,872
TUNE	Focusrite PLC	Consumer Goods	13	10
TUNG	Tungsten Corporation PLC	Financials	10	88
WAND	Wandisco PLC	Technology	691	276
WYG	WYG PLC	Industrials	4	73
YOU	Yougov PLC	Consumer Services	12	2

Table A.3
Main Market (MM) companies (with collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
ACA	Acacia Mining PLC	Basic Materials	3	1518
ADM	Admiral Group PLC	Financials	1239	7
BLT	BHP Billiton PLC	Basic Materials	902	22
BMY	Bloomsbury Publishing PLC	Consumer Services	2420	3
BTG	BTG PLC	Health Care	58,733	132
CNA	Centrica PLC	Utilities	292	2788
DGE	Diageo PLC	Consumer Goods	27	15
GEC	General Electric Company	Industrials	47	0
GMD	Game Digital PLC	Consumer Services	20	518
GSK	Glaxosmithkline PLC	Health Care	1210	1036
IBM	International Business Machines Corporation	Technology	4582	1
KLR	Keller Group PLC	Industrials	8	15
KNM	Konami Holdings Corporation	Consumer Goods	74	0
PMO	Premier Oil PLC	Oil & Gas	92	5870
PRU	Prudential PLC	Financials	553	110
RIO	Rio Tinto PLC	Basic Materials	638	80
RMG	Royal Mail PLC	Industrials	36	2184
SCT	Softcat PLC	Technology	923	97
SDL	SDL PLC	Technology	12	3
SVS	Savills PLC	Financials	7	7
SVT	Severn Trent PLC	Utilities	37	34
TDE	Telefonica Sa	Telecommunications	20	0
TSCO	Tesco PLC	Consumer Services	960	2663
TTA	Total S.A.	Oil & Gas	17	0
VOD	Vodafone Group PLC	Telecommunications	667	843

Table A.4
Main Market (MM) companies (without collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
AVV	Aveva Group PLC	Technology	11	5
BARC	Barclays PLC	Financials	822	1738
BBYB	Balfour Beatty PLC	Industrials	0	0
BFA	BASF SE	Basic Materials	11	0
BP.	BP PLC	Oil & Gas	0	833
BT.A	BT Group PLC	Telecommunications	52	7660
DEB	Debenhams PLC	Consumer Services	755	1109
ECM	Electrocomponents PLC	Industrials	20	3
GNS	Genus PLC	Health Care	7	4
HFD	Halfords Group PLC	Consumer Services	8	62
HSBA	HSBC Holdings PLC	Financials	170	386
KCOM	KCOM Group PLC	Telecommunications	7	46
MRW	Morrison (Wm) Supermarkets PLC	Consumer Services	57	120
OXB	Oxford Biomedica PLC	Health Care	29	914
PDL	Petra Diamonds Limited	Basic Materials	58	568
PSN	Persimmon PLC	Consumer Goods	28	43
RR.	Rolls-Royce Holdings PLC	Industrials	0	375
SGE	Sage Group PLC	Technology	44	17
SHP	Shire PLC	Health Care	1048	759
TYT	Toyota Motor Corporation	Consumer Goods	2	0
UAI	U and I Group PLC	Financials	7	38
USY	Unisys Corporation	Technology	1	0
UU.	United Utilities Group PLC	Utilities	0	101
WG.	Wood Group (John) PLC	Oil & Gas	0	70
ZCC	ZCCM Investments Holdings PLC	Basic Materials	57	0

References

Abdeen, A., Wu, X., Erickson, R., & Fandy, T. (2015). Twitter K-H networks in action : Advancing biomedical literature for drug search. *Journal of Biomedical Informatics, 56*, 157–168. https://doi.org/10.1016/j.jbi.2015.05.015.

Abraham, I., Alonso, O., Kandylas, V., Patel, R., Shelford, S., & Slivkins, A. (2016). How many workers to ask? Adaptive exploration for collecting high quality labels. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval - SIGIR '16* (pp. 473–482). New York, New York, USA: ACM Press. https://doi.org/10.1145/2911451.2911514.

Barnes, P. (2017). Stock market scams, shell companies, penny shares, boiler rooms and cold calling: The UK experience. *International Journal of Law, Crime and Justice, 48*, 50–64. https://doi.org/10.1016/j.ijlcj.2016.11.001.

Bartov, E., Faurel, L., & Mohanram, P. (2017). Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review, 93*(3), accr–51865. https://doi.org/10.2308/accr-51865.

Bentley, P. J., & Lim, S. L. (2017). Fault tolerant fusion of office sensor data using cartesian genetic programming. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. https://doi.org/10.1109/SSCI.2017.8280827.

Sriram, Bharath (2010). *Short text classification in Twitter to improve information filtering*. Columbus, Ohio: The Ohio State University.

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One, 12*(6), 1–17. https://doi.org/10.1371/journal.pone.0177678.

Brown, E. D. (2012). Will twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. In *Proc. of SAIS* (pp. 36–42).

Cheng, W., & Ho, J. (2017). A corpus study of bank financial analyst reports: Semantic fields and metaphors. *International Journal of Business Communication, 54*(3), 258–282. https://doi.org/10.1177/2329488415572790.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., Tesoni, M., & Lillo, F. (2018). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter. *ACM Transactions on the Web*, 1–18 Retrieved from http://arxiv.org/abs/1804.04406, https://tweb.acm.org/.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

Eiselt, A., & Figueroa, A. (2013). A two-step named entity recognizer for open-domain search queries. In *International joint conference on natural language processing* (pp. 829–833). Retrieved from.http://www.chokkan.org/software/crfsuite.

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2018). Big data fusion model for heterogeneous Financial Market Data (FinDF). In *Proceedings of the 2018 Intelligent Systems Conference (IntelliSys)* (pp. 1085–1101).

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 61*, 115–125. https://doi.org/10.1016/j.dss.2014.02.003.

Geron, A. (2017). *Hands-on machine learning with scikit-learn & tensorflow*.

Gorrell, G., Petrak, J., & Bontcheva, K. (2015). Using @Twitter conventions to improve #LOD-based named entity disambiguation. In *European semantic web conference* (pp. 171–186). Cham: Springer. https://doi.org/10.1007/978-3-319-18818-8_11.

Huizinga, T., Ayanso, A., Smoor, M., & Wronski, T. (2017). Exploring insurance and natural disaster tweets using text analytics. *International Journal of Business Analytics, 4*(1), 1–7. https://doi.org/10.1093/rfs/hhv006.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems, 49*(2), 237–253. https://doi.org/10.1007/s10844-017-0458-3.

Jivani, A. G. (2016). A comparative study of stemming algorithms. *International Journal of Computer Technology Applications, 2*(6), 1930–1938.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28–44. https://doi.org/10.1016/j.inffus.2011.08.001.

Matsuda, Koji, Sasaki, Akira, Okazaki, Naoaki, & Inui, Kentaro (2017). Geographical entity annotated corpus of Japanese microblogs. *Journal of Information Processing, 25*(Jan), 121–130. https://doi.org/10.2197/ipsjjip.25.121.

Li, Q., Shah, S., Nourbakhsh, A., Fang, R., & Liu, X. (2017). Fine-grained sentiment analysis on financial microblogs using word vectors built from StockTwits and Twitter (pp. 852–856).

Liu, Y., Cheng, J., Yan, C., Wu, X., & Chen, F. (2015). Research on the Matthews Correlation Coefficients metrics of personalized recommendation algorithm evaluation. *International Journal of Hybrid Information Technology, 8*(1), 163–172.

Mccrum-gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery, 46*, 38–41. https://doi.org/10.1016/j.bjoms.2007.09.002.

Moreno-ortiz, A., & Fernández-cruz, J. (2015). Identifying polarity in financial texts for sentiment analysis : A corpus-based approach, *198*(Clic), 330–338 https://doi.org/10.1016/j.sbspro.2015.07.451 .

Öğüt, H., Mete Doğanay, M., & Aktaş, R. (2009). Detecting stock-price manipulation in an emerging market: The case of Turkey. *Expert Systems with Applications, 36*(9), 11944–11949. https://doi.org/10.1016/j.eswa.2009.03.065.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures, *85*, 62–73 https://doi.org/10.1016/j.dss.2016.02.013.

Perlich, C., Provost, F., Simonoff, J. S., & Stern, L. N. (2003). Tree Induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research, 4*, 211–255.

Pinto, A., Gonçalo Oliveira, H., Alves, A. O., & Oliveira, H. G. (2016). Comparing the performance of different NLP toolkits in formal and social media text. In *5th Symposium on Languages, Applications and Technologies (SLATE'16)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik p. 3:1–3:16 https://doi.org/10.4230/OASIcs.SLATE.2016.3.

Rajesh, N., & Gandy, L. (2016). CashTagNN : Using Sentiment of Tweets with Cash-Tags to Predict Stock Market Prices.

Ramos Carvalho, N., Almeida, J. J., Henriques, P. R., & Varanda, M. J. (2015). From source code identifiers to natural language terms. *The Journal of Systems and Software, 100*, 117–128. https://doi.org/10.1016/j.jss.2014.10.013.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 67*, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.

Spina, D., Gonzalo, J., & Amigó, E. (2013). Discovering filter keywords for company name disambiguation in twitter. *Expert Systems With Applications, 40*(12), 4986–5003. https://doi.org/10.1016/j.eswa.2013.03.001.

Tjong Kim Sang, E., & van den Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal, 3*, 121–134.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research, 2*(Nov), 45–66.

Tripathy, A., & Rath, S. K. (2017). Classification of Sentiment of Reviews using Supervised Machine Learning Techniques. *International Journal of Rough Sets and Data Analysis, 4*(1), 56–74. https://doi.org/10.4018/IJRSDA.2017010104.

Valverde-albacete, F. J., & Pela, C. (2014). 100% Classification Accuracy Considered Harmful : The Normalized Information Transfer Factor Explains the Accuracy Paradox. *PLoS ONE, 9*(1). https://doi.org/10.1371/journal.pone.0084217.

Verma Scholar, M., Professor, A., & Sofat, S. (2014). Techniques to detect spammers in twitter–A survey. *International Journal of Computer Applications, 85*(10), 27–32.

Vilas, A. F., Evans, L., Owda, M., Redondo, R. P. D., & Crockett, K. (2017). Experiment for analysing the impact of financial events on Twitter. In *Algorithms and Architectures for Parallel Processing* (pp. 407–419).

Wood, P. (2015). Automatic and semi-automatic test generation for introductory linguistics courses using natural language processing resources and text corpora. *GSTF Journal on Education (JEd), 3*(1), 1–6 https://doi.org/10.7603/s40 .